Applications of *ab initio* atomistic simulations to biology

2002 J. Phys.: Condens. Matter 14 2957

(http://iopscience.iop.org/0953-8984/14/11/310)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **14** (2002) 2957–2973 PII: S0953-8984(02)28492-1

Applications of *ab initio* atomistic simulations to biology

Matthew D Segall

Camitro (UK) Ltd, UK
and
Cavendish Laboratory, University of Cambridge, UK

Received 3 September 2001
Published 8 March 2002
Online at stacks.iop.org/JPhysCM/14/2957

Abstract

Biological systems provide a particularly challenging set of problems for the application of *ab initio* quantum mechanical simulations. Despite this, these methods are providing insights into biological structures and processes at an atomistic level.

This paper outlines current methods for first-principles modelling of biological systems. Example applications to the cytochrome P450 family of metabolic enzymes, photoreactive rhodopsin proteins and the calculation of NMR chemical shifts are described. Finally, trends in the development of new algorithms and the biological problems to which they will be applied are discussed.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Other papers in this issue demonstrate the utility of *ab initio* simulations in a wide range of physical sciences. In recent years, these methods have made the transition to the biological sciences and are beginning to impact on the understanding of biological processes at a molecular level.

Biological systems typically exhibit a far higher degree of complexity than those studied in the physical sciences. A typical protein may contain several hundred amino acids, thousands of atoms. Most biological processes involve, not a single protein, but complexes of interacting proteins. Added to this, biological reactions occur in solution and solvent molecules often play a critical role in reaction mechanisms. Simulation of a protein, including a full solvation shell, would require tens of thousands of atoms, far beyond the capability of current methods.

Further difficulties are posed by the fact that biological systems are not only wet, but also warm. Therefore, dynamic effects must also be considered. Many biological reactions take place over milliseconds or even seconds. Current *ab initio* molecular dynamics techniques can be applied over picosecond timescales, many orders of magnitude too short to observe these processes.

Despite the size and complexity of biological systems, it is known that changes involving just a single atom can dramatically affect an entire organism, even leading to its death. For example, substitution of a single atom on a molecule can change it from being a highly toxic carcinogen to a life-saving drug. Thus, a complete understanding of biological systems must, in part, include a detailed understanding of the chemical and physical properties of its constituents on an atomistic scale. This can only be achieved by a predictive quantum mechanical modelling technique.

Given the problems outlined above, it is surprising that *ab initio* simulations can be usefully applied to biological systems at all. However, in many cases, carefully defined questions regarding biological systems can be efficiently answered from first principles. Often, the information provided by these studies is not readily available from other sources.

What characteristics make a problem amenable to an *ab initio* study?

- The problem must be structurally well defined.
  Typically, geometries may be obtained from high-resolution crystal structures. However, snapshots from classical molecular dynamics or estimates of structures obtained from proteins of high sequence homology may be used. However, it should be recognized that even the highest-resolution crystal structure provides only an approximate structure as, in a crystal, the molecule is not in its natural environment.
- The mechanism to be investigated must be localized.
  The limited size of system that may be studied from first principles dictates that long-range effects may not be efficiently investigated. Fortunately, proteins are essentially one-dimensional insulators and hence in most cases their electronic states are naturally localized. However, in many systems long-range structural changes can occur, for example on binding of a small molecule to a protein receptor. These processes are not readily accessible to *ab initio* studies at this time.
- A clear question must be posed.
  Ab initio methods are not appropriate to 'look and see' what happens in a biological process. The cost of first-principles simulations implies that they are not able to investigate the large volume of phase space which may be explored by a biological system. A small number of hypothetical mechanisms or structures must be determined *a priori* for study from first principles.

Given the costs and restrictions of using a first-principles approach, what advantages do first-principles approaches offer over other techniques such as classical molecular dynamics or semi-empirical quantum mechanics? Entire proteins, including a full solvation shell, can be simulated using a classical approach. However, the approximation inherent in the use of a classical potential breaks down if bonds are formed or broken. This makes classical methods unsuitable for studying chemical reactions, which require a quantum mechanical approach. Also, in many situations, classical force fields must be parametrized for the specific system being studied, a process which often involves *ab initio* calculations. Semi-empirical quantum mechanical methods such as MNDO and AM1 also rely on parameters which are fitted to experimental results. If the physics or chemistry of the system to be studied differs from those used to parametrize the semi-empirical Hamiltonian, this introduces a source of uncontrolled error. In contrast, *ab initio* methods are parameter free, offering a 'black box' approach to the simulation of biological processes.

The biggest advantage of *ab initio* methods is the ability to perform 'computational experiments' with a high degree of confidence. These can be used to test hypotheses regarding biological structures or mechanisms. Computational experiments differ from their conventional counterparts in the level of control that can be exerted over the

experimental system. A single degree of freedom can be modified within the model system and its effects observed. This is impossible in conventional experiments, as modifications often have unintentional secondary effects that can be impossible to differentiate from the primary effect being investigated. The use of *ab initio* methods in computational experiments minimizes the risk of violating assumptions implicit in the parametrization of empirical or semi-empirical approaches.

The following section of this paper gives a brief description of the methods commonly employed in the study of biological systems from first principles. In section 3, some examples are given of applications to biological problems. Some future trends in methodology and their areas of application are predicted in section 4 and, finally, conclusions are drawn in section 5.

2. Methodology

2.1. Quantum mechanical methods

As discussed above, the study of biological systems typically requires modelling systems containing large numbers of atoms. This prohibits the use of multireference methods such as coupled cluster or CISD due to their computational cost. Hartree–Fock methods have been used in some simulations (see for example [1, 2]). However, questions have been raised regarding the suitability of Hartree–Fock for studying systems containing transition metals [3, 4] (about 30–40% of all functional proteins) and, in particular, hydrogen bonds, which are important in almost all biological systems. Complete active space SCF (CASSCF) methods have been applied in some systems, particularly for spectroscopic applications (e.g. [3] and references therein). However, the most common method applied to biological systems is density functional theory (DFT), due to the good accuracy this approach achieves with high efficiency in large systems. A detailed description of DFT can be found in [5, 6].

Within the Kohn–Sham scheme for DFT [7] the computational cost of calculations scales as the cube of the system size. Although this power is lower than that of correlated wavefunction approaches, this limits the size of a biological system which may be studied to hundreds of atoms on current computers. Linear scaling algorithms for DFT are under development, although these have not been widely applied to date. These will be discussed further in section 4.

Large-scale DFT simulations of biological systems are often performed with a plane wave basis set. Although localized basis sets require fewer basis functions per atom than plane waves, within a Kohn–Sham scheme the computational cost scales linearly with the number of plane waves, while the cost when using a localized basis set scales as the cube of the number of basis functions. Thus, although localized basis sets are widely used, their application is usually limited to systems containing few more than 100 atoms. Plane waves offer further advantages; convergence of results with respect to basis set can be monitored by varying only a single parameter, the cut off energy; there is no basis-set superposition error, as the basis functions are nonlocal; and forces may be efficiently calculated as no Pulay term is required. The last of these points is particularly important. The uncertainty or unavailability of experimental structures of biological systems requires structural relaxations to be performed to identify ground state geometries. Molecular dynamics simulations also require the efficient calculation of forces. However, the advent of linear scaling algorithms, which require localized basis functions, will eventually lead to the dominance of localized basis sets in this field.

Calculations using plane wave basis sets require the use of pseudopotentials, which represent the interaction between valence electrons and the potential due to the nucleus and core electrons by a potential dependent on the atomic species. The most common scheme for construction of pseudopotentials is that of Troullier and Martins [8]. Some plane wave pseudopotential

codes now allow the use of Vanderbilt ultrasoft pseudopotentials [9, 10]. These dramatically reduce the cut-off energy, and hence computational cost, required for a given level of convergence. A full description of the plane wave pseudopotential approach may be found in [5].

The choice of exchange–correlation functional to use within DFT is of great importance. The simple local density approximation (LDA) predicts structural properties in the solid state with surprising accuracy. However, LDA does not give good results for bonding energies and other molecular properties. For this, a gradient corrected functional is required [11]. One commonly used functional for the study of biomolecular systems is that of BLYP [12, 13], which has been shown to give accurate results for hydrogen-bonded systems [6, 14]. The hybrid B3LYP functional combines Slater, Hartree–Fock and Becke exchange with correlation terms due to Lee, Yang and Parr and Vosko, Wilk and Nusair [15]. This has been shown to give better results in molecular systems than a pure density functional approach, although it is prohibitively expensive within the plane wave pseudopotential method.

2.2. Overcoming system size limitations

As discussed above, simulations using conventional *ab initio* methods are limited to systems containing a few hundred atoms. Clearly, it is impossible to study all but the simplest biological systems in their entirety from first principles.

The simplest approach to overcoming the limitations on system size is to isolate the elements of the system which are believed to be most important to the process being studied. This assumption should be tested by systematically enlarging the system and ensuring that the results of the calculations do not change significantly. This can be difficult if the smallest fragment that can be isolated is already near the computational limits of the method being applied. The long-range effects of the surrounding structure can be approximated in this approach by including its electrostatic field in the calculation (see, for example, [16]). The effects of solvation can be approximated by a dielectric continuum, using self-consistent reaction field (SCRF) methods.

A better approximation in many cases is to embed the fragment treated quantum mechanically in a larger system which is described by a classical molecular mechanics approach. These QM/MM approaches can treat entire proteins including solvating water molecules. However, it is still important to ensure that the results of the calculation have converged with respect to the size of the quantum region.

The long-range interactions between the quantum and classical regions are described by an electrostatic potential. The difficulty lies in achieving a good description of the interface between these regions, where covalent bonds may be cut. There are two common approaches to this problem, the link-atom and frozen orbital approaches. The former involves the saturation of valence by substituting hydrogen or halogen atoms or parametrized mono-valent pseudoatoms where bonds have been broken [17]. In the frozen orbital approach a single localized hybrid orbital is included for each QM atom at the interface [18]. A review of these approaches, including a comparison at the semi-empirical level of theory, may be found in [19].

The use of QM/MM methods is growing in popularity, although the intricacies in setting up the interface between the classical and quantum regions means that this is not yet a ‘black box’ technique. Ideally, one would like to treat the entire system quantum mechanically and the use of linear scaling techniques would permit this, as discussed in section 4.

2.3. Overcoming timescale limitations

Most biological reactions take place over timescales inaccessible to *ab initio* molecular dynamics simulations. In order to study the energetics of these reactions and identify transition state structures, it is necessary to explore the potential energy surfaces in a systematic way.

There are many approaches to identifying transition state structures. A first approximation may be achieved using the linear synchronous transit (LST) approach, which maximizes the energy along a linear path connecting the reactant and product of the reaction [20]. This gives an upper bound to the barrier height of the reaction. An improved method, quadratic synchronous transit (QST), maximizes the energy along a parabola in phase space linking the reactant and product. These methods give approximations to the transition state, which can be systematically improved using quasi-Newton or eigenvalue following algorithms [21].

An alternative approach is the intrinsic reaction coordinate approach, which involves optimization of the atomic positions while constraining the value of a reaction coordinate. This process is repeated for a number of values of the reaction coordinate identifying the lowest-energy reaction pathway (see for example [22] and [23]).

These techniques will give estimates of the total energy barrier heights for reactions. However, in many reactions entropic contributions are important in stabilizing the transition state. To estimate free energies requires the statistical sampling of phase space along a reaction coordinate. This can be done by thermodynamic integration, whereby the mean force is integrated along the reaction coordinate using a series of constrained molecular dynamics simulations [24, 25].

3. Applications

The first stage in applying *ab initio* methods to biological problems is the identification of a question of relevance in a biological field. While this may be obvious, it is not straightforward for a physical scientist to enter a biological discipline unguided. Indeed, the straightforward determination of the relevance of a particular protein in a biological process is a current difficulty in molecular biology. For this reason, close collaboration with researchers in the target field is almost essential.

While first-principles simulations may be applied to further understanding in a biological discipline, there are other possible goals for the study of biological systems.

An increasing trend in electronics and materials science is the use of bio-organic materials for technological purposes. Examples of this include the proposed use of DNA for creating electronic circuits [26] and the development of light-emitting organic polymers [27]. A better understanding of the properties of the biological systems with the potential for technological application will aid in the development of these biologically inspired materials.

A wide range of experimental probes are employed in the study of biological systems. Many experiments are interpreted using empirical or semi-empirical methods, from which models of structure or mechanism are deduced. While interpretation of data is often straightforward for small systems, the rapid increase in complexity with system size makes a unique interpretation difficult to achieve in larger systems. An *ab initio* approach to calculating experimental observables would provide a concrete link between experiment and hypothetical structural or mechanistic models, completing the cycle connecting theory and experiment. Examples of techniques amenable to this approach include resonance Raman, EXAFS and optical spectroscopy, ESR and NMR. Resonance Raman spectroscopy has been successfully approached in this way by many groups (see, for example, [28]).

The following sections describe examples of the application of *ab initio* simulations to biological systems. The first is the study by the author and others of the cytochrome P450 superfamily of enzymes, which are important for the metabolism of foreign and endogenous compounds in organisms. The study of the rhodopsin chromophore is described in section 3.2. This protein is responsible for the absorption of photons in the retina, triggering the process of vision, and has potential applications in the field of molecular electronics. Finally, the

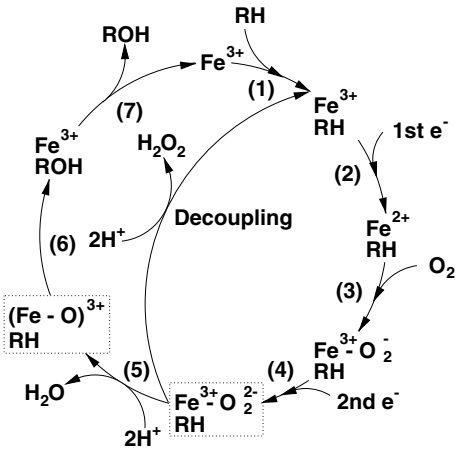

Figure 1. The catalytic cycle of cytochrome P450. RH and ROH represent substrate and hydroxylated substrate respectively.

development of a technique for the calculation of NMR chemical shifts in systems containing hundreds of atoms is described in section 3.3.

3.1. Cytochromes P450

The cytochrome P450 (P450) superfamily of enzymes is present in all classes of organism from bacteria to humans [29]. Members of the family are responsible for many metabolic processes, most importantly the detoxification of a wide range of foreign compounds.

In humans, the P450s present in the gut wall and liver form an important part of the natural defence against foreign compounds entering the body. The wide variety of substrates metabolized by these enzymes presents a challenge in the design of pharmaceutical compounds, which must survive this first-pass metabolic attack and remain in the body for a long period in order to have the desired therapeutic effect. For this reason, P450s have been the target of numerous computational studies, including *ab initio* simulations [30].

The principal reaction catalysed by P450s is hydroxylation of a substrate, which is typically hydrophobic. The main features of the catalytic cycle are shown in figure 1. The reaction occurs at a haem moiety at the active site of the enzyme. This is bound to the backbone of the protein by the sulphur of a cysteine residue. The structure of a bacterial P450, $P450_{cam}$, is shown in figure 2, in which the haem–cysteine system and bound substrate molecule are enlarged. The majority of computational studies have focused on this member of the P450 superfamily as there is an abundance of experimental data available, including high-resolution crystal structures in the free form [31] and in complex with a number of ligand molecules [31–35].

With the exception of the QM/MM study by Murphy *et al* [36], studies of the active site of P450s have involved models of the haem and the ligated cysteine residue in isolation, or including only a small number of surrounding residues. In many studies, the haem–cysteine system was modelled by an iron(III)-S-methyl-porphyrin complex, in which the side chains of the haem were replaced by hydrogens [2, 4, 37]. The use of such a limited model was presumably motivated by the computational cost of including the full structure of the haem; however, the validity of this approximation was not tested. Studies of P450 by the author have included the full haem as shown in figure 2 [38]. The results in table 1 show that the energy difference between the high- and low-spin states, and the charge and spin on the iron, studied in this paper, are not affected by the addition of nearby regions of the protein structure. This confirms that short-range interactions with the surrounding protein do not affect the results. However, the long-range electrostatic field of the protein is not considered in this test.

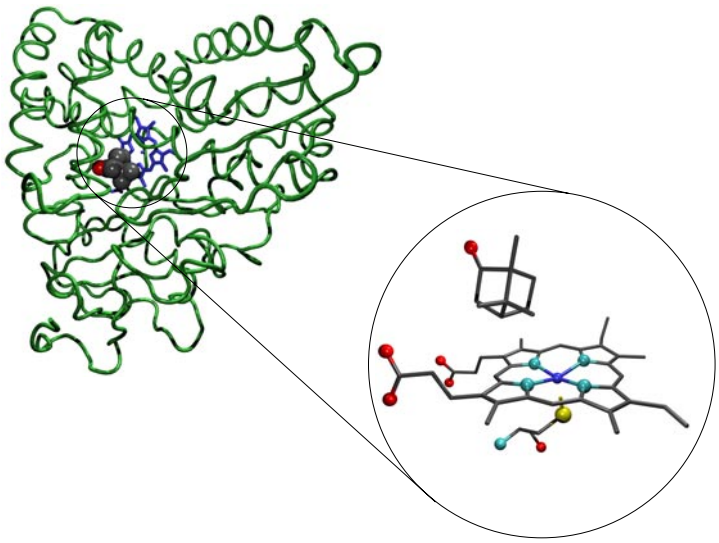

Figure 2. The structure of $P450_{cam}$ with bound camphor [67]. In the protein structure, the backbone is shown in green with the active site haem highlighted in blue and the bound substrate as van der Waals spheres. The active site haem, cysteine 357 residue and bound camphor substrate are shown enlarged. These are the components of the active site studied in [38, 39]. Carbon atoms are shown in grey as bonds only, nitrogen atoms are cyan, iron blue, sulphur yellow and oxygen red.

Table 1. Iron charge and spin, and the energy splitting between the high- (spin $1\hbar$) and low-spin (spin $0\hbar$) states for haem systems including different regions of the protein. These data confirm that the short-range interactions do not significantly change the calculated properties of the P450 model studied. The 'Haem' system includes only the haem moiety, the iron-bound cysteine residue and the camphor substrate.The 'Haem + I helix' system contains the portion of the I helix which crosses the haem on the distal side in addition to the structure in the 'Haem' system. The 'Haem + L' helix system includes residues in the helix on the distal side of the haem, which includes the iron-bound cysteine and the residues which interact most closely with the proprionate side chains of the haem.

System	Fe charge (e)	Fe spin $\hbar$	ΔE_{HS-LS} (eV)
Haem	1.11	1.14	−0.216
Haem + I helix	1.11	1.13	−0.209
Haem + L Helix	1.14	1.15	−0.223

Initial studies of the P450 system focused on the free state of the enzyme, in which a water molecule forms the sixth ligand of the haem iron. In this state, spectrographic studies indicate that the iron is in a low-spin state. On binding of a substrate, step (1) in the catalytic cycle shown in figure 1, the Fe-bound water is removed and the UV absorption spectrum changes dramatically, indicating a change from low to high spin on the iron. Calculations on a series of ligand molecules show an excellent correlation between the calculated and observed spin state of the iron [39]. This gives confidence in the ability of our first-principles techniques to predict the electronic state of the iron.

The next stage of the reaction cycle involves the reduction of the system by an electron transferred from a redox partner. A linear relationship has been found between the redox potential of the system and the spin state of the iron [40]. Calculations of the energy of reduction were found to correlate well with the experimental redox potential and reproduce

Table 2. A comparison of calculated and experimental crystal geometries for the O_2- and CO-bound structures of $P450_{cam}$. Here, the Fe–O1–O2 angle is defined by the haem iron, and the two atoms of the ligand. O1 denotes the closer atom of the ligand (C in the case of CO) and O2 the farther atom of the ligand.

Ligand	Source	Fe–O1–O2 angle (°)	Fe–O1 (Å)	Fe–O2 (Å)
CO	Expt	162	2.0	3.1
CO	Calc.	170	1.8	3.0
O_2	Expt	132	1.8	2.9
O_2	Calc.	123	2.0	3.0

the observed linear relationship with the iron spin [41]. This gives further confidence in the validity of the methods employed.

Having validated the *ab initio* approaches used for the study of the electronic properties of the haem complexes, it is possible to make predictions on intermediate states for which experimental data are not readily available. On binding of molecular oxygen, step (3), the system undergoes rapid reduction (4), followed by formation of the active Fe=O complex (5) or decoupling to form hydrogen peroxide. The nature of these intermediates is therefore not well characterized.

There is conflicting experimental evidence regarding the nature of the singly reduced, oxygen-bound intermediate. Some data indicate that the system is in a ferrous oxide state ($Fe^{2+}O_2$), while others suggest that the system is ferric superoxide ($Fe^{3+}O_2^-$). In contrast to this, CO may bind instead of O_2, forming a stable complex, which has been well characterized experimentally [42]. This system is found to be in a low-spin Fe^{2+} state. *Ab initio* calculations on the CO-bound system find, in good agreement with experiment, the iron is in a low-spin Fe^{2+} state. However, similar calculations indicate that the ground state of the oxygen-bound system is ferric superoxide, resolving the uncertainty regarding this intermediate [41].

Harris and Loew have also studied the oxygen-bound intermediate and obtain broad agreement with the results described here [43]. Furthermore, they studied the protonation of the doubly reduced oxygen-bound intermediate (5). Protonation of the oxygen molecule can lead to the cleavage of the molecular oxygen bond, with the formation of the active Fe=O complex, or decoupling to form hydrogen peroxide, whereby the system returns to the unreduced state. Harris and Loew found that double protonation of the distal atom of the oxygen molecule leads to rapid formation of the active complex, while protonation of the proximal atom followed by protonation of the distal atom leads to a stable protonated state [37]. This indicates that the site of protonation may determine the reaction pathway, to formation of product or hydrogen peroxide.

A crystal structure of the oxygen-bound intermediate of the catalytic cycle has recently been published for $P450_{cam}$ [44]. The crystal structure determined for the oxygen-bound system shows good agreement with the relaxed geometries found in [41]. Table 2 shows a summary of the crystal and calculated geometries for the oxygen- and carbon monoxide-bound systems. The agreement is surprisingly good, as the geometry of the protein structure was fixed during the relaxation and water molecules, subsequently found in the crystal structure, were not included in the calculation.

The crystal structure also shows a potential proton transfer pathway that may be the source for protonation of the molecular oxygen. These structural data may enable the rate of protonation to be predicted and potentially the rate of decoupling for a given substrate. Decoupling to form hydrogen peroxide limits the efficiency of the catalytic reaction and hence

the rate of product formation for many substrates. A better understanding of this process will bring us closer to predicting the rates of metabolism for compounds from first principles.

The careful study of intermediates in the P450 catalytic cycle has improved our understanding of the catalytic mechanism of this important class of drug metabolising enzymes. The next target of these studies is the product formation step of the reaction (6), which will hopefully provide us with the means of predicting the rates and products of P450-mediated drug metabolism from first principles. Techniques have been developed for predicting the products of metabolism, based on the substrate structure alone [45]. However, only a model incorporating the enzyme structure will provide the detailed understanding necessary to make quantitative predictions of absolute reaction rates and regioselectivity.

3.2. Rhodopsin

Rhodopsin proteins consist of a bundle of seven alpha helixes, which support a retinal chromophore. The primary event in the process of vision is the absorption of a photon by the chromophore of a rhodopsin molecule in the retina. The absorption of a photon results in a conformational change in the chromophore, which is followed by a cascade of reactions leading to stimulation of the optic nerve.

In addition to the biological role of rhodopsin proteins, technological applications to molecular electronics are being investigated. In particular, bacteriorhodopsin, an analogue of rhodopsin found in *Halobacterium salinarium* has been suggested as the basis for 'electronic ink', which could be used in inexpensive and low-power electronic displays. This protein changes colour on application of an electric field. An example of a bacterial rhodopsin can be seen in figure 3.

In rhodopsin, the retinal molecule is bound to a lysine residue via a protonated Schiff base. A counterion interacts with the chromophore, which, in bovine rhodopsin, is provided by a glutamate residue [46]. A water molecule hydrogen bonds with both the glutamate and the protonated Schiff base [47]. This system is shown schematically in figure 4. On absorption of a photon, the retinal molecule undergoes isomerization, changing from an 11-*cis* conformation to all-*trans*. This photoisomerization is a very fast process, completing within 200 fs, and has a quantum efficiency of 0.67 [48]. This implies that the process is almost barrierless.

Molteni *et al* studied the energetics of the conformational change of retinal on absorption of a photon [49,50]. This study was undertaken with a DFT-based method. DFT is essentially a ground state methodology; however, a generalization allows study of the potential energy surface in the first excited single state [51].

The initial calculations by Molteni *et al* investigated the energy of the protonated Schiff base of retinal as a function of the C_{10}–C_{11}–C_{12}–C_{13} torsional angle (figure 4 shows the conventional numbering of the carbon atoms in retinal). This was performed in isolation; the counterion and water molecule found in the protein crystal structure were not included in the calculation. In these calculations, a large barrier to rotation was found in both the ground and first excited states. This is not consistent with the observed fast and efficient isomerization.

In contrast, when the counterion and water molecule were included in the simulated system, a dramatic reduction in the height of the barrier to rotation was found. This is consistent with the observation that the speed of isomerization of the chromophore is dramatically reduced in the absence of the protein [52] and demonstrates that inclusion of the interaction between the retinal protonated Schiff base and the protein is essential to understanding the isomerization process. The calculated barrier heights do not agree with those observed experimentally; however, this is probably due to the simple fashion in which the protein environment is included in the model, or limitations in the *ab initio* method used.

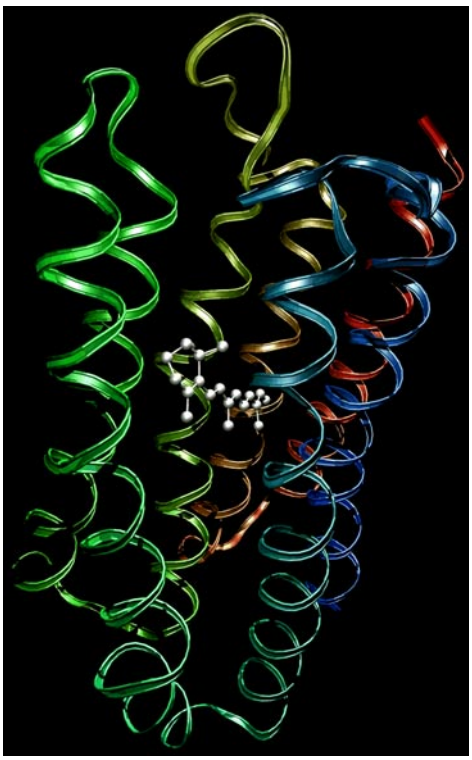

Figure 3. The structure of bacterial sensory rhodopsin II (PDB reference 1JGJ). Rhodopsin proteins consists of seven alpha helixes supporting a retinal chromophore. The backbone of the protein is shaded from blue at the C-termius to red at the N-terminus. The retinal chromophore is shown in grey.

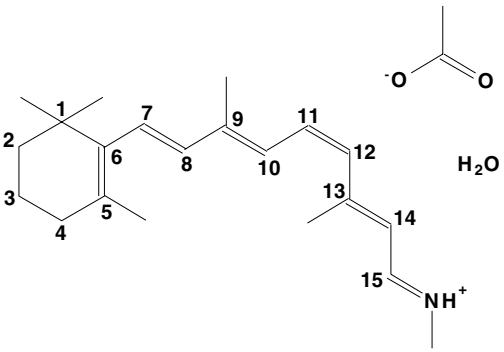

Figure 4. The structure of the protonated Schiff base of retinal, studied by Molteni *et al* in [49,50]. The conventional numbering of the carbon atoms is indicated. Also shown in the figure are a counterion, consisting of the side chain of a glutamate residue, and a water found to form hydrogen bonds with the glutamate and protonated Schiff base.

Further understanding of the effects of the protein environment on isomerization of the chromophore and its absorption spectrum will allow 'tuning' of the properties of rhodopsin proteins for molecular electronics applications.

3.3. NMR chemical shifts

The use of nuclear magnetic resonance (NMR) for structure determination in biology is widespread. Details of this experimental technique can be found in [53].

The NMR technique measures the splitting of the energy levels of nuclei with net spin in an applied magnetic field. The energy level splitting is proportional to the induced magnetic field at the nucleus, which is sensitive to the electronic environment of the nucleus. The chemical shift $\overline{\sigma}$ is defined as the ratio between the induced magnetic field and the external uniform applied field.

$$\boldsymbol{B}_{in}(\boldsymbol{r}) = -\overline{\sigma}(\boldsymbol{r})\boldsymbol{B}. \tag{1}$$

The isotropic chemical shift, measured in solution state NMR, is given by $\mathrm{Tr}[\overline{\sigma}(\boldsymbol{r})]/3$.

Assignment of chemical shifts in a spectrum to specific nuclei, and hence the determination of molecular structure, commonly relies on empirical approaches [53]. These are very powerful for small molecular systems, but for large molecules unique assignments are difficult to deduce. A technique for calculating the chemical shift for each nucleus from first principles, directly from a molecular structure, would provide a concrete link between an observed chemical shift spectrum and a hypothetical structure. This, in turn, would provide a powerful tool for the rapid validation of the results of chemical synthesis and predictions of molecular structure.

First-principles techniques for calculating chemical shifts have been developed using standard quantum chemical approaches [54]. However, the large basis sets needed to achieve convergence of chemical shifts limits these techniques to the study of relatively small molecules (a few tens of atoms). A method recently developed by Mauri and Pickard, based on the plane wave pseudopotential approach [55, 56], extends the range of systems accessible to first-principles chemical shift calculations to include those containing hundreds of atoms. This permits useful applications to biological systems for the first time.

This technique was used by Buda *et al* to study the ^{13}C chemical shifts for retinal systems in rhodopsin, the system discussed in section 3.2. This study found good agreement between the experimental and calculated chemical shifts. Previous analysis of the experimental data, to determine the position of the counterion relative to the retinal chromophore, relied on an empirical linear correlation between the observed ^{13}C NMR chemical shift and the charge density on the conjugated carbon chain [57]. However, the calculations of Buda *et al* showed that the correlation was poor in some charged systems and hence should be used with caution.

Calculations have been recently performed on porphyrin systems, which form the catalytic centres in many biological system (for example P450s discussed above) [58]. Furthermore, arrays of porphyrins can be synthesized to mimic the behaviour of biological systems [59]. NMR is often used to confirm the structure of these porphyrins, both in biological systems and in their synthetic analogues.

Porphyrins are a particularly challenging system to model, due to the wide range of chemical shifts which they exhibit. Porphyrins consist of four pyrole rings (see figure 5 for example) connected by a carbon backbone, which results in an aromatic system in the carbon ring. When an external magnetic field is applied, a 'ring current' is generated, which creates a strong induced field, and hence large chemical shifts. This ring current is a nonlocal effect and presents a challenge to any model for the prediction of chemical shifts.

In [60], Yates *et al* present a study of the porphyrin system shown in figure 5 using the plane wave pseudopotential approach of Pickard and Mauri [56]. Graphs of theoretical predictions versus experimentally observed proton and ^{13}C chemical shifts for this porphyrin are shown in figures 6 and 7 respectively.

First-principles simulations of systems such as porphyrins can improve understanding of experimental observations and their causes. It was previously uncertain whether the ring

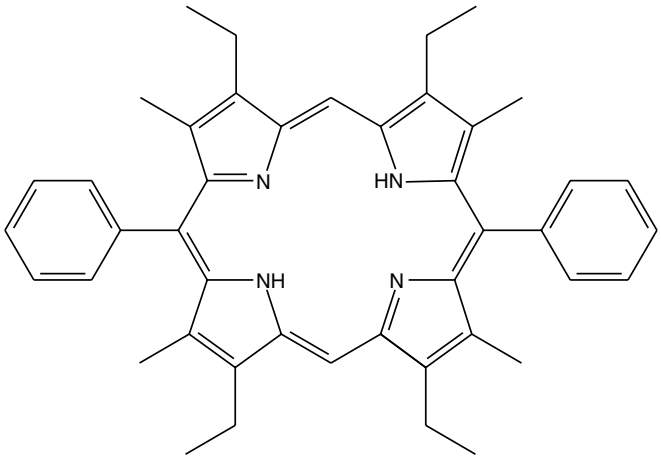

Figure 5. The free-base porphyrin studied in [60]. The main porphyrin ring, formed by four pyrole rings, connected by a carbon backbone, has methyl, ethyl and benzene substitutions. The hydrogens present in the centre of the ring may be replaced by metal ions to form metalloporphyrins such as haem, where an iron atom is substituted.

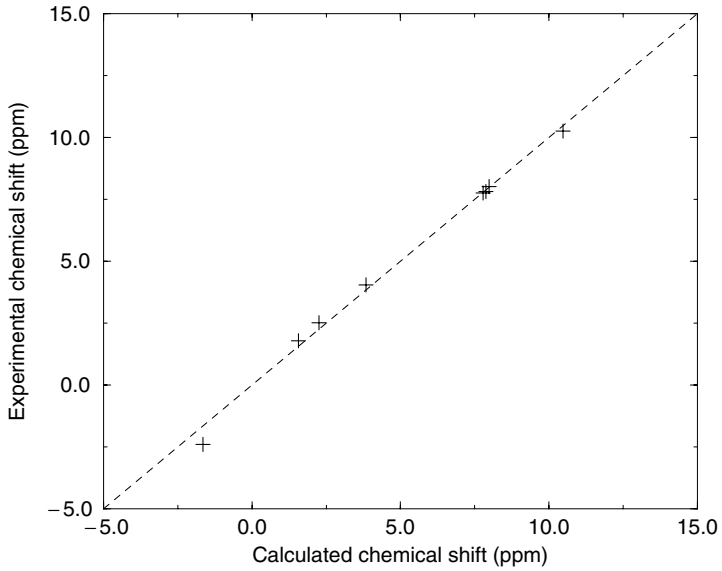

Figure 6. Graph of calculated versus experimental proton chemical shifts for the porphyrin system shown in figure 5. The dashed line indicates the line of perfect agreement.

current was a convenient abstraction for rationalization of experimentally observed chemical shifts or a genuine induced current loop. A plot of the current density in the plane of the porphyrin, shown in figure 8, demonstrates clearly that a genuine current loop results from the application of an external magnetic field.

Currently, chemical shifts may be calculated for diamagnetic systems, i.e. systems with no net spin. The chemical shifts caused by a paramagnetic centre are typically much larger than

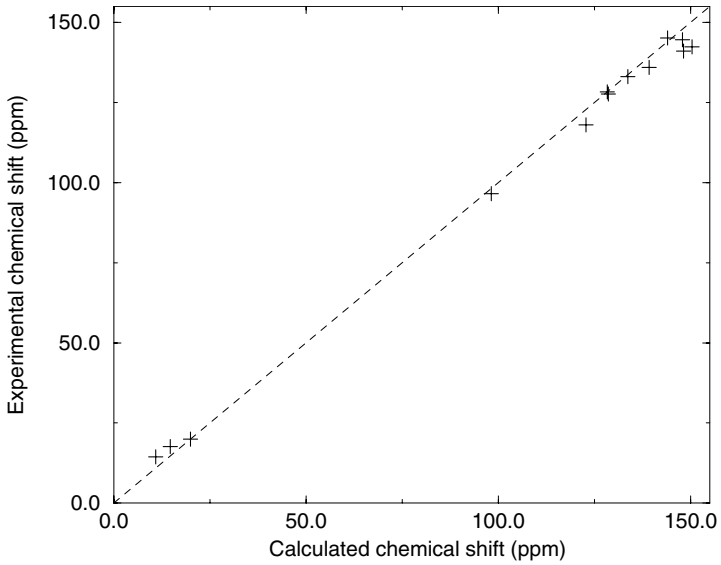

Figure 7. Graph of calculated versus experimental ^{13}C chemical shifts for the porphyrin system shown in figure 5. The dashed line indicates the line of perfect agreement.

those observed in diamagnetic systems. Models for interpretation of these chemical shifts are less mature than those for diamagnetic shifts. Paramagnetic NMR is a rapidly growing field, and can be used, among other things, for rapid structure determination of metalloproteins with a paramagnetic metal ion [61]. The extension of *ab initio* techniques to paramagnetic systems will provide a valuable tool for the interpretation of paramagnetic NMR experiments.

4. Future developments

In the previous sections we have discussed current methodologies for and applications of *ab initio* quantum mechanical simulations. In the future, this field will evolve through the development of new algorithms and their application to new systems. These are discussed in the following two sections.

4.1. New ab initio methods

The most important development is the advent of linear scaling, or O(N), methods for performing *ab initio* calculations. As the name suggests, these scale linearly in computational cost with the size of the system being studied. This compares favourably with conventional techniques, which at best scale as the cube of the system size. This will, for the first time, permit the simulation of entire proteins from first principles.

Traditional cubic scaling methods calculate the single-particle eigenstates and corresponding eigenvalues of the Hamiltonian. If this is performed by direct diagonalization of the Hamiltonian, this procedure scales as N_b^3, where N_b is the number of basis functions. The alternative, iterative diagonalization schemes [5], scale as $N_e^2 N_b$ where N_e is the number of occupied eigenstates, due to the need to maintain orthogonality of the eigenstates during the iterative procedure. The underlying cause of the cubic scaling of these approaches is that the eigenstates extend over the entire system and thus advantage cannot be taken of the inherent locality of quantum mechanics.

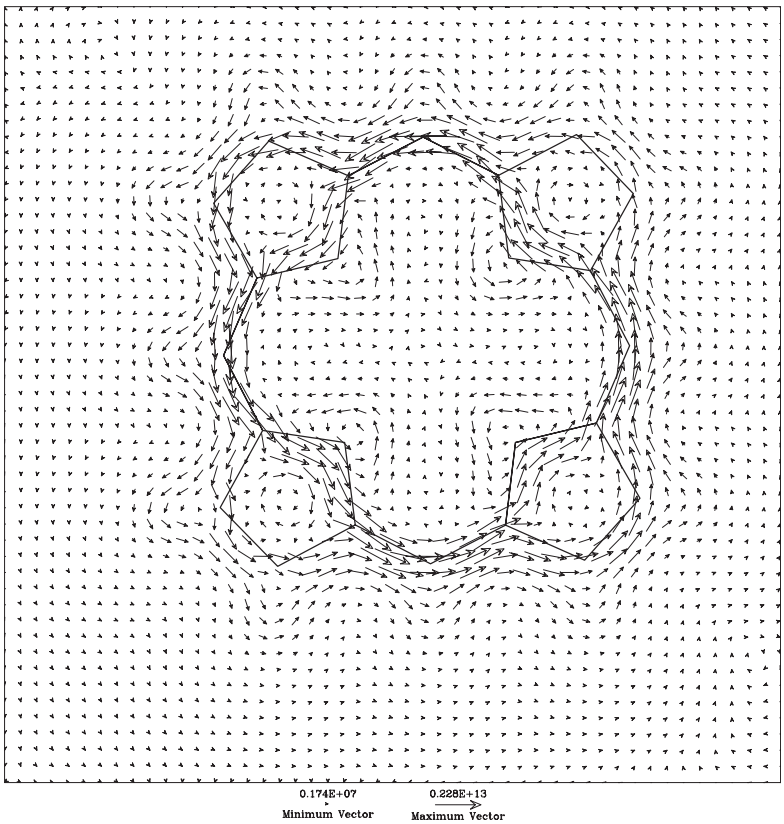

Figure 8. Plot of the current density in the plane of the porphyrin shown in figure 5. The current density is represented by arrows, the direction of the current density in the plane is shown by the direction of the arrow and the magnitude by the length. The structure of the porphyrin ring is superimposed for reference.

In contrast, linear scaling techniques represent the electronic structure in terms of a single-particle density matrix. This is a complete representation of the quantum mechanical system and offers the advantage that the off-diagonal elements of the matrix decay with distance from the diagonal. The exact form of this decay is dependent on the system being studied, but beyond a certain cut-off distance the off-diagonal elements may be neglected. A discussion of linear scaling electronic structure methods may be found in [62].

Recently, applications of *ab initio* O(N) techniques to biological systems have been reported. For example, de Pablo *et al* [63] have performed calculations of the conductivity of λ-DNA, which indicate that λ-DNA chains should have a high resistance, in contrast to some experimental results. The unit cell for these calculations contained 715 atoms and benchmark calculations have been performed on systems containing thousands of atoms [64].

The continued rapid increase in available computational power, coupled with the development of linear scaling techniques, will permit the study of entire proteins from first principles in the near future. However, although such calculations will become tractable, they will remain computationally expensive.

The first routine application of linear-scaling methods may be to validate the assumptions made when studying an isolated fragment of a large biological molecule or embedding such a fragment in a QM/MM simulation. O(N) approaches will enable modelling of sufficiently

large systems from first principles to rigorously test convergence of the results with respect to the size of the quantum mechanically treated fragment.

Ultimately, the study of long-range effects in biological systems with quantum mechanical accuracy will be possible. Such effects include proton and electron transport pathways, rapid long-range conformational changes in response to binding of ligands and interactions of large biological molecules. However, it should be noted that linear scaling methods do not solve the problem of the limited timescale accessible to *ab initio* molecular dynamics.

4.2. New applications

The recently publicized completion of the sequencing of the human genome by the human genome project [65] will yield a wealth of data on the proteins present in the human body. Identification of the approximately 30–45 thousand genes encoded in the 3 billion base pairs of human DNA is well underway. However, this is only the beginning of the process, as determining the function and relevance of these genes and the proteins they encode will present an enormous challenge. The field of functional genomics aims to assign biological functions to genes using a variety of approaches.

The central dogma of molecular biology states that the genetic information encoded in DNA (the genome) is transcribed into RNA (the transcriptome), which in turn controls the synthesis of protein (the proteome) in cell ribosomes. However, in practice, this process is very complex as multiple proteins may result from a single gene by mechanisms such as alternative splicing, whereby sequences of RNA are reordered, or modification of the protein after synthesis. Furthermore, multiple proteins may have the same function or a single protein may fulfill multiple roles.

Statistical and experimental techniques allow the identification of genes related to an observed property (a phenotype) such as susceptibility to a particular disease. In many cases a gene can be translated into the protein sequence it encodes. However, determining the three-dimensional structure of a protein directly from its sequence is not possible at present. Experimental techniques such as x-ray crystallography and NMR rely on obtaining samples of the protein in solution. In most cases this is not straightforward, as the majority of proteins are bound to membranes in the cell and denature if they are removed. Computational techniques such as homology modelling map the sequence of a protein onto the known three-dimensional structure of a protein with similar sequence. These techniques yield approximate three-dimensional geometries which will form the starting point for many *ab initio* simulations in the future.

Cellular functions are regulated by inter-protein interactions or binding of small molecules to enzymes or receptors. Experimental molecular biology approaches can identify the regions of the proteins responsible for these interactions and simulations can yield a more detailed understanding. Mutations in the genes encoding proteins can disrupt important interactions, causing phenotypic variations such as inherited diseases or predisposition to cancer. Understanding these effects will provide the basis for the rational design of drugs that restore or replace the disrupted interactions.

In a similar fashion, the proteins encoded by disease organisms are also important targets for simulations. The basis for drug resistance of bacteria or viruses such as HIV, is mutation of a protein targeted by a pharmaceutical. Detailed understanding of the effects of these mutations will aid the development of treatments for diseases resistant to current treatments. An example of this is the study of HIV-1 reverse transcriptase by Alber and Carloni [16], which investigates a mechanism for resistance to AIDS treatments such as AZT.

A wealth of opportunities for *ab initio* calculations to contribute to biological science will emerge from the genomic revolution. As discussed above, the difficulty will be

identifying targets of genuine biological relevance. For this reason, interdisciplinary collaboration and developing mutual understanding between physical and biological scientists is increasingly important.

5. Conclusion

The application of *ab initio* electronic structure methods to problems in biology is a rapidly growing field. Improvements in algorithms and the availability of increasingly powerful computers have enabled modelling of systems of genuine biological relevance. Techniques for studying complex biological reactions have been developed, enabling their mechanisms to be studied in atomistic detail.

First-principles simulations have been applied to a number of problems in the biological sciences and the use of biological molecules in technological applications. However, the genomic revolution will provide many more questions. Current techniques and those under development will provide powerful tools to aid in the understanding of the biological processes revealed. These insights will provide a basis for future advances in biological and medical sciences. The key to this progress is the continued trend towards closer interdisciplinary links between the biological and physical sciences.

Acknowledgments

The author would like to thank the participants in the CECAM/ESF Ψ_k workshop, '*Ab initio* methods in the biological sciences' held on 11–13 June, 2001, for their stimulating presentations and discussions.

The financial support of Glaxo Wellcome Research and Development, and ML Laboratories plc. is acknowledged for the author's work described in section 3.1. The current support of Camitro (UK) Ltd is gratefully acknowledged.

Figures 2 and 3 were generated with the VMD program. VMD was developed by the Theoretical Biophysics Group in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign [66].

Figure 8 was generously provided by Jonathan Yates.

References

[1] Peeters A and Van Alsenoy C 1999 *Biopolymers* **50** 697
[2] de Groot M J, Havenith R W A, Vinkers H M, Zwaans R, Vermeulen N P E and van Lenthe J H 1998 *J. Comput.-Aided Mol. Des.* **12** 183
[3] Siegbahn P E M and Blomberg M R A 2000 *Chem. Rev.* **100** 421
[4] Green M T 1998 *J. Am. Chem. Soc.* **120** 10 772
[5] Payne M C, Teter M P, Allan D C, Arias T A and Joannopoulos J D 1992 *Rev. Mod. Phys.* **64** 1045
[6] Parrinello M 1997 *Solid State Commun.* **102** 107
[7] Kohn W and Sham L J 1965 *Phys. Rev.* **140** 1133A
[8] Troullier N and Martins J L 1991 *Phys. Rev.* B **43** 1993
[9] Vanderbilt D 1990 *Phys. Rev.* B **41** 7892
[10] Laasonen K, Pasquarello A, Car R, Lee C and Vanderbilt D 1993 *Phys. Rev.* B **47** 10 142
[11] Juan Y M and Kaxiras E 1993 *Phys. Rev.* B **48** 14 944
[12] Becke A D 1988 *Phys. Rev.* A **38** 3098
[13] Lee C, Yang W and Parr R C 1988 *Phys. Rev.* B **37** 785
[14] Molteni C and Parinello M, 1998 *J. Am. Chem. Soc.* **120** 2168
[15] Becke A D 1993 *J. Phys. Chem.* **98** 5648
[16] Alber F and Carloni P 2000 *Protein Sci.* **9** 2535

[17] Eichinger M, Tavan P, Hutter J and Parrinello M 1999 *J. Chem. Phys.* **110** 10 452
[18] Murphy R B, Philipp D M and Friesner R A 2000 *Chem. Phys. Lett.* **321** 113
[19] Reuter N, Dejaegere A, Maigret B and Karplus M 2000 *J. Phys. Chem.* A **104** 1720
[20] Halgren T A and Liscomb W N 1977 *Chem. Phys. Lett.* **49** 225
[21] Peng C Y and Schlegel H B 1993 *Isr. J. Chem.* **33** 449
[22] Peterson T H and Carpenter B K 1992 *J. Am. Chem. Soc.* **114** 766
[23] Sandre E, Payne M C and Gale J D 1998 *Chem. Comm.* 2445–6
[24] Beveridge D L and DiCapua F M 1989 *Annu. Rev. Biophys. Biophys. Chem.* **18** 431
[25] Sprik M and Ciccotti G 1998 *J. Chem. Phys.* **109** 7737
[26] Dekker C and Ratner M 2001 *Phys. World* **14**
[27] Greenham N C, Moratti S C, Bradley D D C, Friend R H and Holmes A B 1993 *Nature* **365** 628
[28] Putrino A, Sebastiani D and Parrinello M 2000 *J. Chem. Phys.* **113** 7102
[29] Ortiz de Montellano P R 1996 *Cytochrome P450. Structure, Mechanism and Biochemistry* 2nd edn (New York: Plenum)
[30] Loew G H and Harris D L 2000 *Chem. Rev.* **100** 407
[31] Poulos T L, Finzel B C and Howard A J 1986 *Biochemistry* **25** 5314
[32] Poulos T L, Finzel B C and Howard A J 1987 *J. Mol. Biol.* **195** 687
[33] Raag R and Poulos T L 1989 *Biochemistry* **28** 7586
[34] Raag R and Poulos T L 1991 *Biochemistry* **30** 2674
[35] Raag R, Li H, Jones B C and Poulos T L 1993 *Biochemistry* **32** 4571
[36] Murphy R B, Philipp D M and Freisner R A 2000 *J. Comput. Chem.* **21** 1442
[37] Harris D L and Loew G H 1998 *J. Am. Chem. Soc.* **120** 8941
[38] Segall M D, Payne M C, Ellis S W, Tucker G T and Boyes R 1998 *Xenobiotica* **28** 15
[39] Segall M D, Payne M C, Ellis S W, Tucker G T and Boyes R N 1998 *Phys. Rev.* E **57** 4618
[40] Fisher M and Sligar S 1985 *J. Am. Chem. Soc.* **107** 5018
[41] Segall M D, Payne M C, Ellis S W, Tucker G T and Eddershaw P J 1998 *Xenobiotica* **29** 561
[42] Lewis D F V 1996 *Cytochromes P450: Structure, Function and Mechanism* (London: Taylor and Francis)
[43] Harris D, Loew G and Waskell L 1998 *J. Am. Chem. Soc.* **120** 4308
[44] Schlichting I *et al* 2000 *Science* **287** 1615
[45] Korzekwa K R, Grogan J, DeVito S and Jones J P 1996 *Biological Reactive Intermediates* vol 5, ed R Snyder (New York: Plenum) pp 361–9 ch 44
[46] Jäger F, Fahmy K and Sakmar T P 1994 *Biochemistry* **33** 10 878
[47] Deng H, Huang L W, Callender R and Ebrey T 1994 *Biophys. J.* **66** 1129
[48] Schoenlein R W, Peteanu L A, Mathies R A and Shank C V 1991 *Science* **254** 412
[49] Molteni C, Frank I and Parrinello M 1999 *J. Am. Chem. Soc.* **121** 12 177
[50] Molteni C, Frank I and Parinello M 2001 *Comput. Mater. Sci.* **20** 311
[51] Frank I, Hutter J, Marx D and Parrinello M 1998 *J. Chem. Phys.* **108** 4060
[52] Logunov S L, Song L and El-Sayed M A 1996 *J. Phys. Chem.* **100** 18 586
[53] Evans J N S 1995 *Biomolecular NMR Spectroscopy* (Oxford: Oxford University Press)
[54] Cheeseman J R, Trucks G W, Keith T A and Frisch M J 1996 *J. Chem. Phys.* **104** 5497
[55] Mauri F, Pfrommer B G and Louie S G 1996 *Phys. Rev. Lett.* **77** 5300
[56] Pickard C J and Mauri F 2001 *Phys. Rev.* B **6324** 5101
[57] Han M, Dedecker B S and Smith S O 1993 *Biophys. J.* **65** 899
[58] Milgrom L R 1997 *The Colours of Life: an Introduction to the Chemistry of Porphyrins and Related Compounds* (Oxford: Oxford University Press)
[59] Mak C C, Bampos N, Darling S L, Montalti M, Prodi L and Sanders J K M 2001 *J. Org. Chem.* **66** 4476
[60] Yates J R, Pickard J C, Mauri F, Segall M D and Payne M C in preparation
[61] Bertini I and Luchinat C 1996 *Coordination Chemistry Reviews* vol 150, ed A B P Lever (Amsterdam: Elsevier)
[62] Goedecker S 1999 *Rev. Mod. Phys.* **71** 1085
[63] de Pablo P J, Moreno-Herrero F, Colchero J, Herrero J G, Herrero P, Baro A M, Ordejon P, Soler J M and Artacho E 2000 *Phys. Rev. Lett.* **85** 4992
[64] Bowler D R and Gillan M J 2000 *Mol. Simulat.* **25** 239
[65] http://www.ornl.gov/hgmis/
[66] Humphrey W, Dalke A and Schulten K 1996 *J. Mol. Graphics* **14** 33
[67] Raag R, Martinis S A, Sligar S G and Poulos T L 1991 *Biochemistry* **33** 11 420